Analyzing Consumer Mindset Metrics in Arabic Dialectical Texts on Social Media Platforms Using Deep Learning

Safa Al-Sarayreh
School of Computer Sciences
Universiti Sains Malaysia, Penang
Malaysia
safaalsarayreh@student.usm.my

Mohd Husin
School of Computer Sciences
Universiti Sains Malaysia, Penang
Malaysia
heikal@usm.my

Noor Ibrahim
School of Computer Sciences
Universiti Sains Malaysia, Penang
Malaysia
nfarizah@usm.my

Abstract: *In particular, the Arabic text on social media provides a wealth of information about consumer sentiment, attitudes, and behavior, as digital change occurs more quickly than ever. Examining this form of text is difficult because of the morphological complexity of the Arabic language and its dialect variability. Deep learning models were employed to address these challenges to classify Arabic social media comments into satisfaction, loyalty, purchase intention, and service quality. Modern deep learning techniques, such as Arabic Bidirectional Encoder Representations from Transformers (AraBERT) and Bidirectional Long Short-Term Memory Network (BiLSTM), are used in this research. The results demonstrate the importance of the models employed, with AraBERT yielding the best results across all measurements concerning the unbalanced dataset. Meanwhile, when it comes to a balanced dataset, the BiLSTM performs slightly better than AraBERT. This research provides evidence to support the viability of these models for the classification of short Arabic text. It lays a foundation for applying the deep learning model to derive insights from Jordanian dialect's social media comments. The impact of the balanced dataset on the performance of deep learning models is also confirmed by this research.*

Keywords: *Deep learning, consumer mindset, arabic dialectical texts, social media platforms, consumer behavior, NLP.*

Received June 1, 2025; accepted October 12, 2025
https://doi.org/10.34028/iajit/23/1/3

1. Introduction

Classifying Arabic texts is crucial for effectively processing and evaluating Arabic-language content in a variety of digital formats, including subject modeling, sentiment analysis, and information retrieval [7]. It is considered a key area of study for machine learning and natural language processing [4]. It is defined as the automatic assignment of the most appropriate labels or categories to textual content [24]. Because the Arabic language is a highly complex language, with a multitude of dialects and roots of complexity [14, 23]. It might be quite challenging to categorize customer feedback in Arabic [42].

Arabic text classification has recently been boosted by the development of Natural Language Processing (NLP) resources, but the task is still challenging because of the nature of the Arabic language. Yet, texts in Arabic are still reported to be very difficult compared to many other languages [8]. Such challenges the language presents include rich morphology, orthographic ambiguity, dialectal variation, orthographic noise, and a lack of annotated resources for Machine Learning (ML) and Deep Learning (DL) model training and evaluation [22]. In several NLP tasks, including text classification tasks, large pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) achieve state-of-the-art performance [26]. Moreover, Arabic text classification is difficult because of the language's complex morphology, the dearth of publicly accessible corpora, and the fact that Modern Standard Arabic (MSA) exists along with several dialects [16]. Similarly, more research is needed to classify Arabic text, including resolving the issue of unbalanced datasets and improving the preprocessing stage. It has been demonstrated that methods based on deep learning are more effective than traditional machine learning techniques [9]. The Arabic Bidirectional Encoder Representations from Transformers (AraBERT) model is selected due to its extensive training on Arabic text and ability to be used in NLP applications, including question answering, text categorization, and language translation [2].

In addition, Bidirectional Long Short-Term Memory Network (BiLSTM), an extension of Long Short-Term Memory Network (LSTM), can process both forward and backward text, allowing it to capture contextual information better than regular LSTM models [32]. Thus, the primary benefit of BiLSTM is a two-way structure, which allows for richer and more accurate text representation by capturing context from both preceding and subsequent words [25].

This research contributes by developing an Arabic consumer review dataset in the Jordanian dialect and

analyzing customer reviews through perceptual constructs that define consumer behavior. Our method provides companies with more in-depth and valuable customer information than previous research, which mostly relies on sentiment analysis.

Therefore, this research employs deep learning to classify Arabic comments collected from multiple Jordanian retail pages on Facebook and Instagram. The collected comments are manually labeled according to the four Consumer Mindset Metrics (CMMs): satisfaction, loyalty, purchase intent, and service quality.

This research is organized as follows: section 2 presents the relevant literature on Arabic text classification and consumer mindset metrics. Section 3 presents the research methodology, including data collection, annotation, preprocessing, model construction, and results. Section 4 concludes, outlining the main conclusions, limitations, and possible directions for further investigation.

2. Literature Review

This section first examines previous research on CMM, which serves as the conceptual foundation for this research. Then, it goes over text classification techniques in Arabic, emphasizing their applicability and difficulties for social media comment analysis. This gives readers a thorough understanding of the research context.

2.1. Consumer Mindset Metrics

Consumer metrics include variables such as customer satisfaction, service quality, loyalty, and purchase intention, which are defined as endogenous perceptual variables [27]. These metrics are the primary perceptual constructs that will dominate this research.

Consumer behavior reflects customers' attitudes toward products or services. Since corresponding attitude-related measures aim to answer the why question required to interpret observed consumer behavior, they are also known as CMMs [31]. Marketing knowledge reveals that mindset metrics have been in use, mainly within branding and advertising arenas, to measure consumers' state of mind, perceptions, and behaviors [35]. Also, CMMs record how customers feel about their experiences with a business, good, or service [27, 41, 45]. Studying CMMs in marketing models is so crucial [40, 48]. Therefore, they are key influencers of sales and corporate value; it is crucial to understand their origins [17, 48].

2.2. Text Classification (TC)

Text classification is the process of labeling texts according to pre-defined categories [24]. Research on Arabic Text Categorization (ATC) has recently gained more attention, especially when social media data are assessed and customers' attitudes are understood. Another challenge that scholars encounter when considering the classification of tweets is that the short text may not be sufficient to provide adequate context [17]. This issue is similar to the comments on Instagram and Facebook. Furthermore, these difficulties are more significant in classifying CMMs, which are crucial for understanding customer behavior in the Jordanian retail industry, including satisfaction, loyalty, service quality, and purchase intention.

In a different study Zaghloul *et al*. [51], examined how retailers can accurately estimate customer satisfaction in e-commerce to make better decisions using machine learning, such as Support Vector Machine (SVM) and Logistic Regression (LR). Moreover, Mozumder *et al*. [37] uses a framework that combines ML and DL with BERT to categorize customer feedback into different customer satisfaction drivers, including the American Customer Satisfaction Index (ACSI) and the Net Promoter Score (NPS). The fine-tuned BERT model was trained on 5943 customer feedback responses from 39 companies across 13 industries, and the model had an F1-score of 0.84.4. This result outperformed the baseline approaches, such as the SVM, with an F1-score of 0.47. Moreover, AraBERT was used to categorize Saudi telecom tweets, demonstrating that transformer-based Arabic models outperformed Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) baselines in generalizing and capturing consumer satisfaction, with an accuracy of approximately 94% [2]. In addition, a comparison of machine learning, deep learning, and pre-trained models on multilingual e-commerce corpora revealed that they all perform differently. The Bert-base-multilingual-cased model outperformed the BiLSTM and SVM baselines and achieved the highest accuracy on the Arabic sentiment analysis (95.1%) [43]. Similarly, the BERT-LSTNet-Softmax model outperformed BiLSTM and other baselines in e-commerce review analysis, achieving up to ~97% accuracy and demonstrating its excellent efficacy in predicting sentiment and purchase intention [34].

According to Alshamari [11], a DL approach is employed to study users' satisfaction with the services offered by the Saudi Telecom Company (STC), Mobily, and Zain, particularly on sentiment analysis of social media reviews. DL models, including LSTM, Gated Recurrent Unit (GRU), and BiLSTM, were used to analyze the publicly available AraCust dataset. LSTM demonstrated the maximum accuracy, training accuracy of 98.04%, and test accuracy of 97.03%. Hence, the study emphasizes the importance of customer satisfaction measurement in telecommunications since the quality of services determines customer choices and loyalty.

Stop-word removal, tokenization, lemmatization, or similar normalization techniques, as well as Term Frequency-Inverse Document Frequency (TF-IDF), are some of the most typical preprocessing and feature extraction methods identified in the literature [20]. Several studies have faced unbalanced datasets. Hence, a balanced dataset is required to enhance learning and classification when using machine learning for classification. Because the categories in the dataset being collected are imbalanced and unequal, this problem has not been extensively discussed in the ATC literature [49]. This is considered one of the most significant problems in text classification [6]. Literature analysis identifies a data gap: the possibility of creating incredibly complex deep learning models is limited due to the nature of the Arabic language and the lack of training datasets. Arabic label text classification is crucial and challenging due to the scarcity of large, open-source Arabic-rich datasets.

Some research questions include handling language ambiguity, creating open data sets, and developing models to classify consumer mindset metrics (satisfaction, loyalty, purchase intent, and service quality). These developments will make efficient and productive machine-learning applications in ATC possible.

Based on the literature reviewed, this methodology will focus on an elaborate approach to Arabic text classification. The preprocessing steps are tokenization, stop word removal, and feature extraction employing TF-IDF. Besides, the study employs different deep learning models like BiLSTM and AraBERT. This approach is designed to develop a strong model specific to classifying social media comments in Jordan. Despite its significance, Arabic text classification is still underexplored. A summary of Arabic text classification has been studied and is presented in Table 1.

Table 1. A review of Arabic text classification research, focusing on developing approaches in the area.

Reference	Dataset	Classifier	Performance	Train-test split	Preprocessing techniques	Classes
[9]	SANAD, HARD, AJGT, ASTD, ArsenTD-Lev	BERT model	Accuracy: 97.89%, 96.78%, 92.10%, 85.25%, 52.25%	80:20	Removes Arabic stop words, punctuation, and extra spaces	-
[29]	From Facebook, Instagram, and X (Twitter)	LSTM, LDA, KNN	Best: LSTM (Acc 85%)	80:20	Tokenization, word encoding	Health, Security, Services, Organization, Worship
[44]	Arabic poem-emotion (9452 poems)	Deep learning (DCNN, GRU, LSTM), AraBERT	AraBERT model: Accuracy 76.5%, F1-score improvement by 24%	80:20	Removal of prepositions, tokenization, rooting, stemming, feature extraction	Sadness, Joy, Love
[5]	AraSenCorpus (15,000 tweets, 34 million extended)	LSTM	F1-score: 87.4%	-	Tokenization, normalization	Positive, Negative, Neutral
[13]	3,000 notes from various telecommunication companies, in the Jordanian dialect	SVM and BiLSTM	66% 99.33%	Training, validation, and testing as 80%, 10%, and 10%.	Stemming Stop word removal Non-Arabic removal	Positive, Negative, Neutral
[11]	AraCust	LSTM, GRU, and BiLSTM	97.03% 96.82% 96.40%	80:20	Tokenization various preprocessing strategies	Satisfaction dissatisfaction
[10]	10,646 various coffees products' Twitter reviews	K-nearest neighbor, support vector machine, decision tree, and random forest	74.0% 95.0% 95.0% 94.0%	80:20	Pre-processing TF-IDF	Positive, Negative, Neutral
[36]	(ANP5) from Arabic News Posts from several Arabic platforms	RF SVM LR NB BIGRU CNN-LSTM LSTM CNN	82.05 82.0% 81.0% 79.0% 88.10% 89.30% 89.85% 90.10%	Training, validation, and testing as 80%, 10%, and 10%.	Preprocessing approaches remove @date and @time symbols, punctuation marks, diacritics, strip elongation, normalization, and stopwords	Positive Negative
[30]	7,000 Arabic	Random forest, Decision tree, K-neighbors, Gradient Boosting, and XGBoost	93.0% 94.0% 86.0% 71.0% 86.0%	10-fold cross-validation	Pre-processing WordNet	Spam Ham

3. Methodology

This section describes the research's methodological framework. This section begins with a data collection procedure and then discusses preprocessing, annotation, and model building. Each stage ensures the strength and dependability of the Arabic social media comment classification procedure.

3.1. Data Collection

In this research. The initial dataset for this study consisted of customer reviews from Jordanian retail industries, written in an Arabic Jordanian dialect, including 6,710 comments gathered from widely used social media retail Jordanian pages, particularly Facebook and Instagram. The retail shops that provided these comments were Al_Bayrouty, Ameed Coffee, Crispy Chicken, Durra Markets, Lafamilia, Sereensallalboutique, and Vikik Fashion. As part of our research analysis, the dataset was used as the basis for implementing and assessing various deep learning methods. The comments are in Arabic, Jordanian dialect. Moreover, because the dataset is imbalanced. We oversampled the minority classes by randomly duplicating samples using the resample function from sklearn.utils until all classes had the same number of

samples as the majority class. Table 2 shows the label counts of the dataset before and after.

Table 2. Labels count before and after random oversampling.

Label	Unbalanced	Balanced
Satisfaction	3109	3109
Loyalty	1599	3109
Purchase intent	1315	3109
Service quality	687	3109

The random oversampling method is selected because of its effectiveness, ease of use, and improved performance on unbalanced datasets [50]. By adding minority class samples, random oversampling balances the dataset and improves classifier performance [19]. This method uses balanced data to train the Arabic text classification model, reducing bias toward the majority class and producing more objective predictions [18].

Random oversampling ensures that the model is not biased toward the majority class during training. The dataset is divided into 80% training and 20% testing data [17]. Furthermore, by reproducing instances of minority classes, random oversampling reduces bias toward majority classes and improves the fairness of text categorization algorithms [18].

Figure 1 depicts the overall study design, using Arabert and BiLSTM for classification.

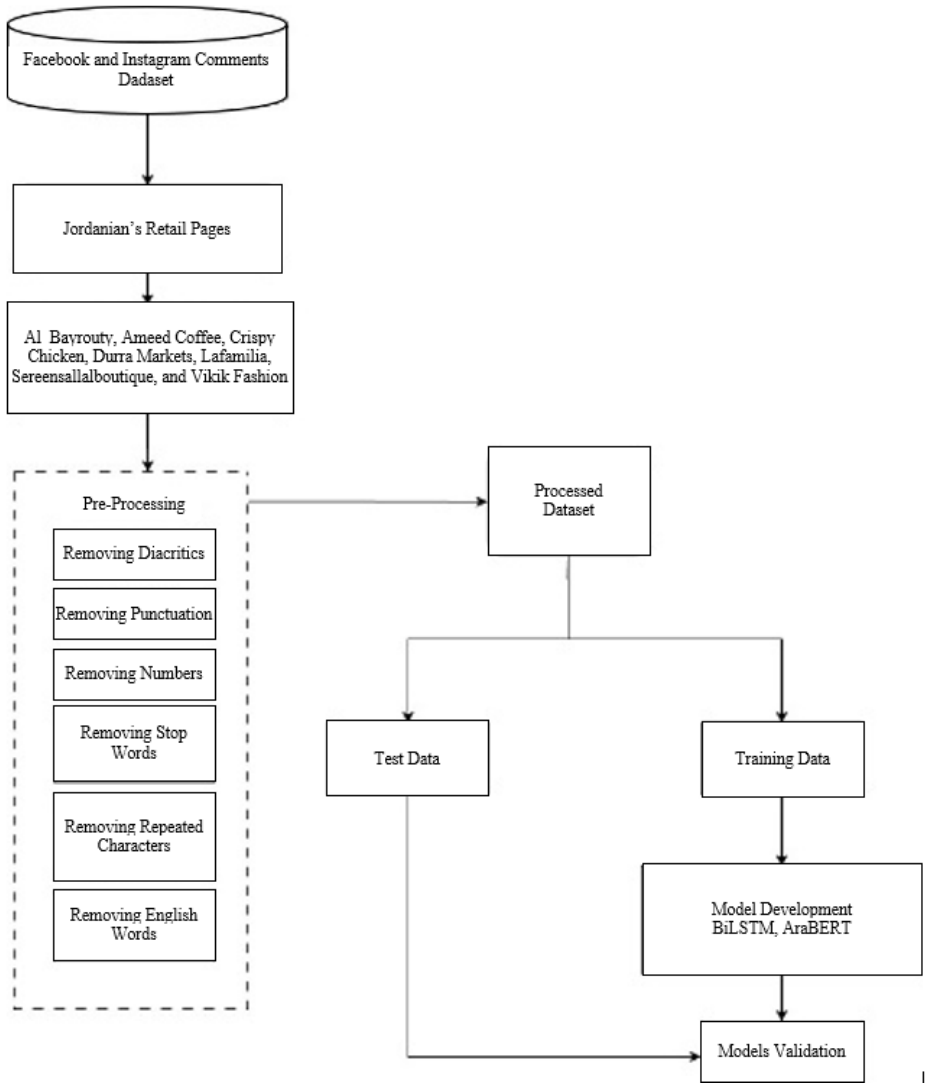

Figure 1. The proposed methodology for the Arabic text classification task.

3.2. Data Pre-Processing

Text pre-processing is an essential step in cleaning the dataset and enhancing the results [24]. Several procedures are carried out during the data cleaning to increase the models' accuracy. Eliminates additional spaces, non-Arabic characters, numbers, punctuation, diacritical marks, Tatweel, and Arabic letters are normalized (e.g., converting different forms of alef "أ , إ , آ ," to " ا ", and " ى " to " ي "), and removing URLs. The Arabic list provided by NLTK is used to eliminate stop words. The Arabic Snowball stemmer is then used to stem the text. Also, removing Arabic stop words is performed while retaining essential negation words such as "لن", "ما", "ليس", "لا", and "لم". With a maximum sequence length of 128 characters, the texts are tokenized using the AraBERT v2 tokenizer. For BiLSTM, the NLTK Arabic tokenizer and Snowball stemmer are used. Table 3 shows the preprocessing steps applied to the dataset.

Table 3. Pre-processing steps applied to the dataset.

Preprocessing Steps	Example before processing	Example after processing	Translation
Step 1: Removing punctuation	كم سعر البلوزه وسايز واحد؟؟	كم سعر البلوزه وسايز واحد	How much is the blouse? Is there only one size?
Step 2: Removing numbers	رقم ١ ❤❤❤👐👐👐	رقم 👐👐❤❤❤	Number 1
Step 3: Removing diacritics	ما شاءالله	ما شاء الله	Praise be to God
Step 4: Removing english words	Banan Al-Ababneh bana بجننو صح	بجننو صح	So beautiful
Step 5: Removing repeated characters	كتيييير حلو	كتير حلو	Very nice
Step 6: Removing stop words	هاد المحل اللي بجيب منه الجكيتات Rawan Nobani	هاد المحل بجيب الجكيتات Rawan Nobani	This is the store I get jackets from
Step 7: Removing mentions (@user)	@Noor Ayed شو رايك نروح😍	شو رايك نروح😍	What do you think, shall we go?

3.3. Data Annotation

Annotating datasets is the primary difficulty in preparing a training dataset for machine learning tasks like text categorization. In this Research, data annotation was conducted manually in MS Excel sheets by native Arabic speakers proficient in consumer mindset metrics and the Jordanian Arabic dialect. During the annotation phase, three annotators who are familiar with the CMM perceptual (satisfaction, loyalty, service quality, and purchase intent). The dataset is labeled by reviewing each comment and assigning a label based on CMMs.

In particular, Vargas *et al*. [47] calculate inter-annotator agreement using the three metrics: Cohen's kappa, Fleiss' kappa, and simple inter-annotator agreement. Besides, when examining agreements between more than two raters and assessing nominal categories, Fleiss' kappa is utilized [21]. Whereas, a value of 1 indicates perfect agreement, and 0 indicates random agreement [28]. The annotators for this study annotate each comment as follows:

- Satisfaction indicates whether or not the customer is satisfied with the brand.
- Loyalty indicates if the customer is committed to the same brand.
- Service quality indicates the degree to which a service satisfies or meets the customer's expectations.
- Purchase intent indicates the customer's intention to buy the product/service.

3.4. Model Construction and Training

Our research selected models based on Arabert and BiLSTM classifiers as the baseline models to classify Jordan's retail industries' social media comments into four categories: satisfaction, loyalty, purchase intent, and service quality. 80% of the data is used for training, and 20% for testing. A classification model is trained using a training set, which consists of known examples. Subsequently, the model could forecast unknown samples with the most likely class, often accomplished by evaluating performance against a testing set. In addition, classification is labeling data documents or text into their classes based on their content [38]. In the training phase, the classifier algorithm gains knowledge from the labeled data, which in our case consisted of customer reviews within a Jordanian retail setting. As a result, the classifier has to be able to categorize new and unlabeled customer feedback during the testing phase. Next, we measured each classifier's accuracy, F-score, precision, and recall. By calculating the area under the ROC curve, which shows the trade-off between true-positive and false-positive rates, Area Under the Curve (AUC) measures a classifier's overall efficacy [12].

3.4.1. AraBERT

Researchers and practitioners are paying growing attention to BERT since it is a beneficial method for processing natural languages, trained exclusively for Arabic, and uses 8.2 billion tokens of text from Wikipedia and other Arabic sites in Modern Standard Arabic [3]. Nonetheless, the multilingual version received further training on 10 million tweets in various Arabic dialects [15]. A trained language model for Arabic language processing tasks is called AraBERT. The fact that AraBERT was trained on a sizable dataset of Arabic text indicates that it has a solid command of the language and is capable of high-accuracy language translation, text categorization, and sentiment analysis, among other possible benefits.

The AraBERT v2 tokenizer (aubmindlab/bert-base-arabertv02) is used in the present study to extract features. The tokenizer transforms each preprocessed Arabic comment into a series of subword tokens and maps them to dense vector embeddings in a fixed-length vector space. By capturing semantic and contextual information, these embeddings enable the model to depict the meaning of individual words and their connections within the phrase. Three epochs are used.

3.4.2. Bidirectional Long Short-Term Memory Network (BiLSTM)

BiLSTM is an expansion of the LSTM paradigm and is made up of two LSTMs that run in opposite directions [32]. Contextual dependencies and local semantic features are both captured by the BiLSTM layer, while the attention layer highlights the most informative elements associated with each label [33].

The Arabic Snowball stemmer stems the text once it has been tokenized using NLTK. Every comment is converted into a fixed-length sequence of token IDs. A trainable 100-dimensional embedding feeds a 2-layer BiLSTM (hidden size 128, dropout 0.3) that learns dense feature representations during training. The BiLSTM model trains for 10 epochs.

3.5. Evaluation Metrics

Employing different feature representation methods, the classifiers' performance is evaluated on both balanced and unbalanced versions of the dataset. The performance of the proposed technique is assessed using the following metrics: measures such as F-Measure, recall, accuracy, and precision have been identified as fit for measuring the effectiveness of the proposed models. Accuracy is the proportion of correctly classified labels, whereas recall is the proportion of actual positive labels that are correctly classified.

The research employs the following six metrics to compare performance [39]:

- The correct prediction percentage measures classification accuracy, as shown in Equation (1), where *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative.

$$Accuracy = (TP + TN)/(FP + TP + FN + TN) \quad (1)$$

- The percentage of genuine positives among cases anticipated to be positive is known as precision, as shown in Equation (2).

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

- The percentage of true positives among all positive data instances is known as recall, as shown in Equation (3).

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

- A weighted harmonic mean of recall and precision is called F1, as shown in Equation (4).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

- ROC: is a graphical method for examining a classifier's performance [46].
- AUC: is an additional metric for assessing quality that establishes a correlation between true positives and false positives. From this perspective, an algorithm is considered better the higher its AUC value [1].

3.6. Experimental Analysis and Discussion

In addition to selecting the optimal classification, performance evaluation involves confirming that the classification satisfies design specifications and pointing out areas where the classification component needs to be improved. Arabic comments are manually classified and put through a series of preprocessing steps, including tokenization, stemming, filtering, and feature extraction. The Arabic comments are then turned into numerical data called vector features, which are subsequently fed into deep Learning, such as Arabert and BiLSTM, building models. The calculated Fleiss' Kappa value of 0.995 reflects an exceptionally high level of agreement among the raters annotating the dataset. To tackle this text classification issue in the first study, a dataset of 6710 comments was first created. Table 4 shows a sample of the dataset used. Table 5 shows comments with rater-assigned labels and label count distribution.

Table 4. Sample of the dataset.

Comments	Label	Translation
شو رايك نروح😊😍	Purchase intent	What do you think, shall we go? 😊😍
السعر	Purchase intent	The price
كم سعره وين موقعكم	Purchase intent	How much is it? Where is your location?
احسن لحوم من الدره نشتري كل مره اي شي زاكي	Loyalty	The best meat from Al-Durra, we buy something delicious every time.
انا دايما بشتري اللحمه من عندكن شكرا لكم بالتوفيق ان شاء الله	Loyalty	I always buy meat from you. Thank you, and best of luck, God willing.
كتير حلو	Satisfaction	Very nice
نرجس تشرين بجنن بدي واحد	Satisfaction	Nargis Tishreen is amazing, I want one.
لازمنا مشوار	Purchase intent	We need a trip.
هاد المحل اللي بجيب منه الجكيتات	Loyalty	This is the store where I buy jackets.
والله يوم الجمعة جربنها وشي فاخر وكرم الضيافة كان حاضر مع معاملات الشباب كتير راقية وممتازة وصرة ملبس من المولد النبوي الشريف وشرحات على عجين ممتازة	Service quality	We tried it on Friday, and it was luxurious. The hospitality was great, and the staff was very professional and excellent. The sweets from the Prophet's birthday celebration and the meat slices on dough were excellent.
الأسعار ارتفعت بشكل كبير بالنرويج والسويد. يرجى المتابعه والنظر بجديه للأسعار. هناك من يستغل الوضع..	Service quality	Prices have risen significantly in Norway and Sweden... Please follow up and seriously consider the prices. Some people are taking advantage of the situation.
بس ياريت الانتقال بين الصور ما يكون سريع .. ما عم نلحق نقرأ العروض ولا حتى نوقف الفيديو عشان نقرأ الانتقال بين الصور سريع جدا	Service quality	I wish the transition between images wasn't so fast… We can't keep up with reading the offers or even pause the video to read. The transition is too quick.

Table 5. Comments with rater-assigned labels and label count distribution.

Comments	Purchase intent	Loyalty	Satisfaction	Service quality	Predominant category
في زيت حلو واديش سعر ليتر	3	0	0	0	Purchase intent
متألقين دايما اختياري الصح انتو❤	0	3	0	0	Loyalty
بجننو صح	0	0	3	0	Satisfaction
شوفي حلو	0	0	3	0	Satisfaction
حلوين صح ☹	0	0	3	0	Satisfaction

We discuss how dataset class imbalance affects the evaluation of deep learning classifiers. As shown in Table 6, the AraBERT model's results applied to the unbalanced dataset show how well it can categorize Arabic social media text, with high assessment metrics overall. The model performs exceptionally well in

classification with an AUC of 99.1% and an accuracy of 94.2%. With an F1-score of 94.2%. According to the confusion matrix, the model shows a high accuracy and balanced classification performance across the labels, with small misclassifications across related labels, particularly between loyalty and satisfaction, even if most classes are correctly categorized. These findings imply that AraBERT can manage the complexity of Arabic text and hold promise for precisely categorizing consumer mindset metrics. This result also demonstrates how well AraBERT handles the complexities of the Arabic language. Comparing these results to the metrics of the balanced dataset, it is asserted that AraBERT achieves a slightly higher performance with an accuracy of 94.5%. With 94.5% for the F1-score and 99.2% for the AUC.

The results of the BiLSTM model evaluation show strong performance across key metrics. For the unbalanced dataset, the model demonstrates an accuracy of 92.7%. F1-score, Precision, and recall are all 92.7%, indicating a good balance. According to this, the model detects true positives while maintaining a low number of false positives and false negatives across classes, which is crucial for unbalanced multi-class classification. Furthermore, the overall AUC is 98.9%, indicating strong class separability across satisfaction, loyalty, purchase intent, and service quality, and providing dependable insights into consumer mindset metrics with minimal errors. In addition, the diagonal in the confusion matrix shows the correctly classified labels. Their high values demonstrate strong per-class accuracy. However, there is a clear enhancement regarding the balanced dataset, with an accuracy of 94.6%, and 94.6% for F1, precision, and recall.

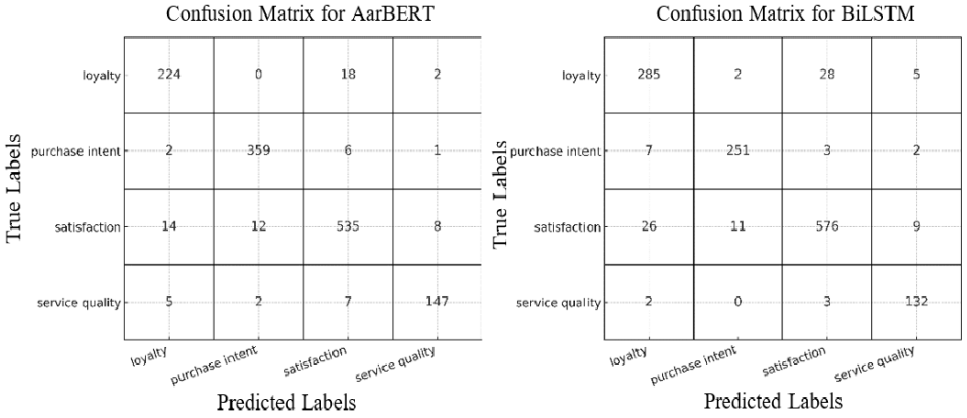

Figure 2. Confusion matrix for the AraBERT and BiLSTM models for the unbalanced dataset.

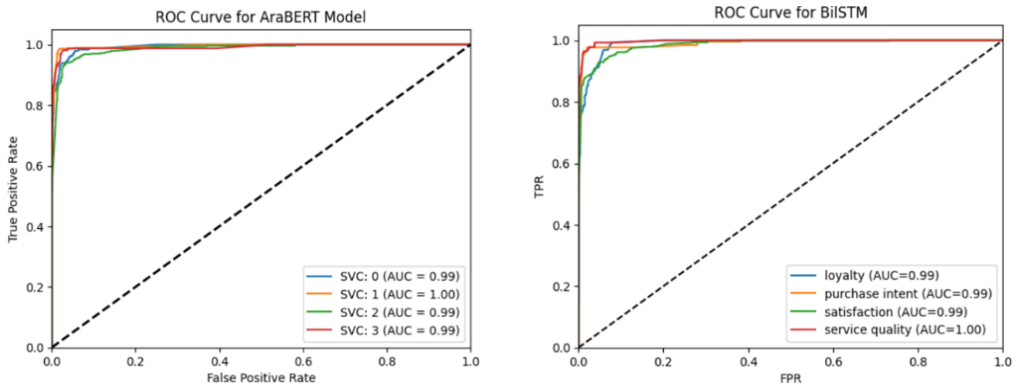

Figure 3. ROC curve for AraBERT and BiLSTM model for unbalanced dataset.

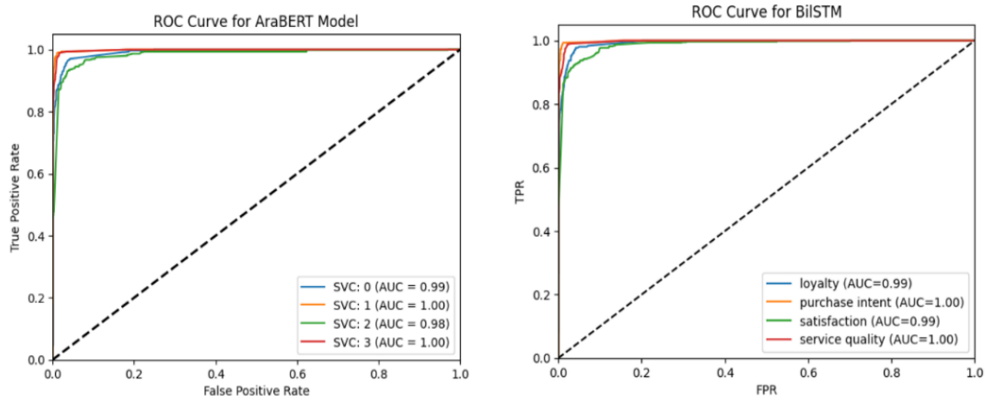

Figure 4. ROC curve for AraBERT and BiLSTM model for balanced dataset.

Figure 5. Confusion matrix for the AraBERT and BiLSTM models for the balanced dataset.

Figure 4 shows the ROC curve for the AraBERT and BiLSTM models for the balanced dataset. When applying the balanced dataset, it is noticed that there is an enhancement in the performance of both AraBERT and BiLSTM, as shown in Table 4. Furthermore, Figure 5 shows the Confusion Matrix for the AraBERT and BiLSTM models for the Balanced Dataset. Figure 2 shows the confusion matrix for the AraBERT and BiLSTM models for the unbalanced dataset. Figure 3 also shows the ROC curve for the AraBERT and BiLSTM models for the unbalanced dataset. Regarding Arabic text multi-class categorization, the AraBERT model achieves up to 99% accuracy and state-of-the-art performance results [47].

Table 6 displays the outcomes of the deep learning algorithms. Figure 5 shows the confusion matrix of the balanced dataset.

A comparison of deep learning techniques based on the F1-Score value, precision, recall, and test data accuracy is displayed in Figure 6.

Table 6. Deep learning algorithm results.

	Unbalanced dataset					Balanced dataset				
Model	**F1-Score (%)**	**Accuracy (%)**	**AUC (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Accuracy (%)**	**AUC (%)**	**Precision (%)**	**Recall (%)**
AraBERT	94.2	94.2	99.1	94.2	94.2	94.5	94.5	99.2	94.5	94.5
BiLSTM	92.7	92.7	98.9	92.7	92.7	94.6	94.6	99.4	94.6	94.6

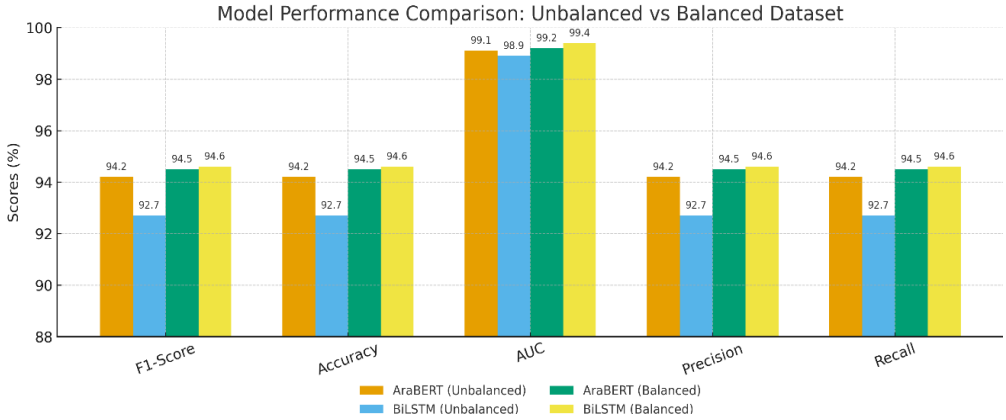

Figure 6. Evaluation metrics among DL algorithms comparison.

4. Conclusions

This research uses a primary dataset gathered from multiple Jordanian retail pages on Facebook and Instagram, together with deep learning models like Arabert and BiLSTM, to offer an automated categorization technique for consumer mindset metrics. AraBERT performs better than BiLSTM in terms of an unbalanced dataset, with an accuracy of 94.2%. This approach reflects Arabic's contextual richness and is extremely useful, particularly for grasping the language's subtleties. Likewise, BiLSTM yields good results, confirming the efficacy of deep learning architectures for Arabic text, with an accuracy of 92.7%. In addition, the BiLSTM slightly outperforms the AraBERT for a balanced dataset with 94.6% accuracy. Similarly, the ability to generalize text classification to other dialects or languages may be limited since the linguistic features may vary considerably, which could impact the model's functionality [13]. Thus, we recommend investigating multi-Arabic datasets for CMM classification. Moreover, this research relies only on AraBERT and BiLSTM models. Future research should consider additional deep learning classifiers and hybrid models to enhance classification performance and address the challenges posed by Arabic's morphological richness.

In conclusion, deep learning models show higher skills in Arabic text categorization and provide consistency and computational efficiency. Besides,

BiLSTM shows promising results for a balanced dataset. These findings conclude the significance of the distribution of the classes on deep learning models. This comparative analysis highlights the effectiveness of utilizing sophisticated NLP frameworks for languages like Arabic, where capturing context and subtleties is crucial. Furthermore, the research's conclusions also provide Jordanian retailers with valuable recommendations on identifying customer satisfaction, loyalty, service quality, and purchase intent.

References

[1] Abd D., Khan W., Khan B., Alharbe N., and et al., "Categorization of Arabic Posts Using Artificial Neural Network and Hash Features," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 6, pp. 102733, 2023. DOI: 10.1016/j.jksus.2023.102733

[2] Aftan S. and Shah H., "Using the AraBERT Model for Customer Satisfaction Classification of Telecom Sectors in Saudi Arabia," *Brain Sciences*, vol. 13, no. 1, pp. 1-20, 2023. DOI: 10.3390/brainsci13010147

[3] Alammary A., "BERT Models for Arabic Text Classification: A Systematic Review," *Applied Sciences*, vol. 12, no. 11, pp. 5720, 2022. DOI: 10.3390/app12115720

[4] Alhawarat M. and Aseeri A., "A Superior Arabic Text Categorization Deep Model," *IEEE Access*, vol. 8, pp. 24653-24661, 2020. https://doi.org/10.1109/ACCESS.2020.2970504

[5] Al-Laith A., Shahbaz M., Alaskar H., and Rehmat A., "Arasencorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus," *Applied Sciences*, vol. 11, no. 5, pp. 2434, 2021. DOI: 10.3390/app11052434

[6] Almuzaini H. and Azmi A., "Impact of Stemming and Word Embedding on Deep Learning-based Arabic Text Categorization," *IEEE Access*, vol. 8, pp. 127913-127928, 2020. DOI: 10.1109/ACCESS.2020.3009217

[7] Alnagi E., Ghnemat R., and Abu Al-Haija Q., "Boosting Arabic Text Classification Using Hybrid Deep Learning Approach," *Discover Applied Sciences*, vol. 7, no. 540, 2025. https://doi.org/10.1007/s42452-025-07025-x

[8] Al-Qerem A., Raja M., Taqatqa S., and Abu Sara M., "Utilizing Deep Learning Models (RNN, LSTM, CNN-LSTM, and Bi-LSTM) for Arabic Text Classification," *Artificial Intelligence-Augmented Digital Twins*, vol. 503, pp. 287-301, 2024. DOI: 10.1007/978-3-031-43490-7_22

[9] Alruily M., Fazal A., Mostafa A., and Ezz M., "Automated Arabic Long-Tweet Classification Using Transfer Learning with BERT," *Applied Sciences*, vol. 13, no. 6, pp. 1-17, 2023. DOI: 10.3390/app13063482

[10] Alsemaree O., Alam A., Gill S., and Uhlig S., "Sentiment Analysis of Arabic Social Media Texts: A Machine Learning Approach to Deciphering Customer Perceptions," *Heliyon*, vol. 10, no. 9, pp. e27863, 2024. https://www.cell.com/heliyon/fulltext/S2405-8440(24)03894-5

[11] Alshamari M., "Evaluating User Satisfaction Using Deep-Learning-based Sentiment Analysis for Social Media Data in Saudi Arabia's Telecommunication Sector," *Computers*, vol. 12, no. 9, pp. 1-24, 2023. DOI: 10.3390/computers12090170

[12] Al-Smadi B., "DeBERTa-BiLSTM: A Multi-Label Classification Model of Arabic Medical Questions Using Pre-Trained Models and Deep Learning," *Computers in Biology and Medicine*, vol. 170, pp. 107921, 2024, DOI: 10.1016/j.compbiomed.2024.107921

[13] Alsokkar A., Otair M., Alfar H., Nasereddin A., and et al., "Sentiment Analysis for Arabic Call Center Notes Using Machine Learning Techniques," *Journal of Autonomous Intelligence*, vol. 7, no. 3, pp. 1-16, 2024. DOI: 10.32629/jai.v7i3.940

[14] Alsuwaylimi A., "Arabic Dialect Identification in Social Media: A Hybrid Model with Transformer Models and BiLSTM," *Heliyon*, vol. 10, no. 17, pp. 1-18, 2024. DOI: 10.1016/j.heliyon.2024.e36280

[15] Al-Twairesh N., "The Evolution of Language Models Applied to Emotion Analysis of Arabic Tweets," *Information*, vol. 12, no. 2, pp. 1-15, 2021. DOI: 10.3390/info12020084

[16] Alwehaibi A., Bikdash M., Albogmi M., and Roy K., "A Study of the Performance of Embedding Methods for Arabic Short-Text Sentiment Analysis Using Deep Learning Approaches," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6140-6149, 2022. DOI: 10.1016/j.jksuci.2021.07.011

[17] Alzanin S., Azmi A., and Aboalsamh H., "Short Text Classification for Arabic Social Media Tweets," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6595-6604, 2022. DOI: 10.1016/j.jksuci.2022.03.020

[18] Angdresey A., Sitanayah L., Lucky I., and Tangka H., "Sentiment Analysis for Political Debates on YouTube Comments Using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes," *Journal of Computing Theories and Applications*, vol. 2, no. 3, pp. 342-354, 2025. https://doi.org/10.62411/jcta.11668

[19] Ashisha G., Mary X., Kanaga E., Andrew J., and Eunice R., "Random Oversampling-based Diabetes Classification via Machine Learning Algorithms," *International Journal of*

Computational Intelligence Systems, vol. 17, no. 270, pp. 1-17, 2024. DOI: 10.1007/s44196-024-00678-3

[20] Ashokkumar P., Arunkumar N., and Don S., "Intelligent Optimal Route Recommendation Among Heterogeneous Objects with Keywords," *Computers and Electrical Engineering*, vol. 68, pp. 526-535, 2018. DOI: 10.1016/j.compeleceng.2018.05.004

[21] Bartok L. and Burzler M., "How to Assess Rater Rankings? A Theoretical and a Simulation Approach Using the Sum of the Pairwise Absolute Row Differences," *Journal of Statistical Theory and Practice*, vol. 14, no. 37, pp. 1-16, 2020. DOI: 10.1007/s42519-020-00103-w

[22] Bourahouat G., Abourezq M., and Daoudi N., "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal of Information Technology* , vol. 21, no. 2, pp. 313-325, 2024. DOI: 10.34028/21/2/13

[23] El Rifai H., Al Qadi L., and Elnagar A., "Arabic Text Classification: The Need for Multi-Labeling Systems," *Neural Computing Applied*, vol. 34, no. 2, pp. 1135-1159, 2022. https://doi.org/10.1007/s00521-021-06390-z

[24] Elnagar A., Al-Debsi R., and Einea O., "Arabic Text Classification Using Deep Learning Models," *Information Processing and Management*, vol. 57, no. 1, pp. 102121, 2020. https://doi.org/10.1016/j.ipm.2019.102121

[25] Fitriyah Z. and Kartikasari M., "Text Classification of Twitter Opinion Related to Permendikbud 30/2021 Using Bidirectional LSTM," *Barekeng Jurnal Ilmu Matematika Dan TerapaN*, vol. 17, no. 2, pp. 1113-1122, 2023. DOI: 10.30598/barekengvol17iss2pp1113-1122

[26] Galal O., Abdel-Gawad A., and Farouk M., "Rethinking of BERT Sentence Embedding for Text Classification," *Neural Computing Applied*, vol. 36, no. 32, pp. 20245-20258, 2024. DOI: 10.1007/s00521-024-10212-3

[27] Gupta S. and Zeithaml V., "Customer Metrics and their Impact on Financial Performance," *Marketing Science*, vol. 25, no. 6, pp. 718-739, 2006. DOI: 10.1287/mksc.1060.0221

[28] Kankeviciute E., Songailaite M., Mandravickaite J., Kalinauskaite D., and Krilavicius T., "A Comparison of Deep Learning Models for Hate Speech Detection," *in Proceedings of the 27th International Conference on Information Technology*, Kaunas, pp. 1-11, 2022. https://ceur-ws.org/Vol-3611/paper19.pdf

[29] Khan M. and AlGhamdi M., "A Customized Deep Learning-based Framework for Classification and Analysis of Social Media Posts to Enhance the Hajj and Umrah Services," *Expert Systems with Applications*, vol. 238, pp. 122204, 2024. DOI: 10.1016/j.eswa.2023.122204

[30] Kraidia I., Ghenai A., and Belhaouari S., "A Multi-Faceted Approach to Trending Topic Attack Detection Using Semantic Similarity and Large-Scale Datasets," *IEEE Access*, vol. 13, pp. 21005-21028, 2025. DOI: 10.1109/ACCESS.2025.3535996

[31] Kübler R., Adler S., Welke L., and Sarstedt M., "Mining Consumer Mindset Metrics with User-Generated Content," *Schmalenbach Journal of Business Research*, vol. 77, pp. 497-525, 2025. DOI: 10.1007/s41471-025-00219-4

[32] Liu C., "Long Short-Term Memory (LSTM)-Based News Classification Model," *PLoS One*, vol. 19, no. 5, pp. 1-23, 2024. DOI: 10.1371/journal.pone.0301835

[33] Lu G., Liu Y., Wang J., and Wu H., "CNN-BiLSTM-Attention: A Multi-Label Neural Classifier for Short Texts with a Small Set of Labels," *Information Processing and Management*, vol. 60, no. 3, pp. 103320, 2023. DOI: 10.1016/j.ipm.2023.103320

[34] Ma X., Li Y., and Asif M., "E-Commerce Review Sentiment Analysis and Purchase Intention Prediction Based on Deep Learning Technology," *Journal of Organizational and End User Computing*, vol. 36, no. 1, pp. 1-29, 2024. DOI: 10.4018/JOEUC.335122

[35] Marshall P., "A Latent Allocation Model for Brand Awareness and Mindset Metrics," *International Journal of Market Research*, vol. 64, no. 4, pp. 526-540, 2021. DOI: 10.1177/14707853211040052

[36] Mhamed M., Sutcliffe R., and Feng J., "Benchmark Arabic News Posts and Analyzes Arabic Sentiment Through RMuBERT and SSL with AMCFFL Technique," *Egyption Informatics Journal*, vol. 29, pp. 100601, 2025. DOI: 10.1016/j.eij.2024.100601

[37] Mozumder A., Nguyen T., Devi S., Ahmed P., and et al., "Enhancing Customer Satisfaction Analysis Using Advanced Machine Learning Techniques in Fintech Industry," *Journal of Computer Science and Technology Studies*, vol. 6, no. 3, pp. 35-41, 2024. DOI: 10.32996/jcsts.2024.6.3.4

[38] Muaad A., Davanagere H., Guru D., Benifa J., and et al., "Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques," *Mathematical Problems in Engineering*, vol. 2022, no. 1, pp. 1-16, 2022. DOI: 10.1155/2022/3720358

[39] Oyebode O., Alqahtani F., and Orji R., "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps based on User Reviews," *IEEE Access*, vol. 8, pp. 111141-111158, 2020. DOI: 10.1109/ACCESS.2020.3002176

[40] Petersen J., Kumar V., Polo Y., and Sese F., "Unlocking the Power of Marketing: Understanding the Links between Customer Mindset Metrics, Behavior, and Profitability," *Journal of the Academy of Marketing* Science, vol. 46, no. 5, pp. 813-836, 2018. DOI: 10.1007/s11747-017-0554-5

[41] Rubera G. and Kirca A., "You Gotta Serve Somebody: The Effects of Firm Innovation on Customer Satisfaction and Firm Value," *Journal of the Academy of Marketing Science*, vol. 45, no. 5, pp. 741-761, 2017. DOI: 10.1007/s11747-016-0512-7

[42] Salloum A. and Almustafa M., "Analysis and Classification of Customer Reviews in Arabic Using Machine Learning and Deep Learning," *Journal of Data Acquisition and Processing*, vol. 38, no. 4, pp. 726-744, DOI: 10.5281/zenodo.777803

[43] Savci P. and Das B., "Prediction of the Customers' Interests Using Sentiment Analysis in E-Commerce Data for Comparison of Arabic, English, and Turkish languages," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 3, pp. 227-237, 2023. DOI: 10.1016/j.jksuci.2023.02.017

[44] Shahriar S., Al Roken N., and Zualkernan I., "Classification of Arabic Poetry Emotions Using Deep Learning," *Computers*, vol. 12, no. 5, pp. 1-14, 2023. DOI: 10.3390/computers12050089

[45] Srinivasan S., Vanhuele M., and Pauwels K., "Mind-Set Metrics in Market Response Models: An Integrative Approach," *Journal of Marketing Research*, vol. 47, no. 4, pp. 672-684, 2010. DOI: 10.1509/jmkr.47.4.672

[46] Tan P., *Receiver Operating Characteristic*, Encyclopedia of Database Systems, 10.1007/978-0-387-39940-9_569, Last Visited, 2025.

[47] Vargas F., Carvalho I., Goes F., Benevenuto F., and Pardo T., "Building an Expert Annotated Corpus of Brazilian Instagram Comments for Hate Speech and Offensive Language Detection," *in Proceedings of the 13th Conference on Language Resources and Evaluation*, Marseille, pp. 7174-7183, 2021. https://aclanthology.org/2022.lrec-1.777/

[48] Venkatesan R., Bleier A., Reinartz W., and Ravishanker N., "Improving Customer Profit Predictions with Customer Mindset Metrics Through Multiple Overimputation," *Journal of the Academy of Marketing Science*, vol. 47, no. 5, pp. 771-794, 2019. DOI: 10.1007/s11747-019-00658-6

[49] Wahdan A., Al-Emran M., and Shaalan K., "A Systematic Review of Arabic Text Classification: Areas, Applications, and Future Directions," *Soft Computing*, vol. 28, no. 2, pp. 1545-1566, 2024. DOI: 10.1007/s00500-023-08384-6

[50] Wongvorachan T., He S., and Bulut O., "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Informaion*, vol. 14, no. 1, pp. 54, 2023. DOI: 10.3390/info14010054

[51] Zaghloul M., Barakat S., and Rezk A., "Predicting E-Commerce Customer Satisfaction: Traditional Machine Learning vs. Deep Learning Approaches," *Journal of Retailing and Consumer Services*, vol. 79, pp. 103865, 2024. DOI: 10.1016/j.jretconser.2024.103865

Safa Al-Sarayreh is currently a postgraduate PhD Student at the School of Computer Sciences, Universiti Sains Malaysia (USM), specializing in natural language processing. She received her Master's degree in Quality Management from the University of Jordan and her Bachelor's degree in Computer Engineering from Princess Sumaya University for Technology in 2007.

Mohd Husin is an IS Senior Lecturer in the School of Computer Sciences at the Universiti Sains Malaysia. He holds a PhD in IT from the University of South Australia. He is also an ISTQB Certified Software Tester and IREB Certified Requirements Engineering. His current research interests include Technology Adoption in organizations, Software Testing, Digital Transformation, and more recently ICT Education.

Noor Ibrahim works as a Senior Lecturer at the School of Computer Sciences, USM after receiving her PhD from the University of Bristol, United Kingdom. She is also a certified tester of the International Software Testing Qualifications Board (ISTQB) and a Huawei Certified ICT Associate and Academy Instructor in Artificial Intelligence (AI). Her research interests include Data/Text Analytics, Natural Language Processing, Machine Learning, and Social Media Research. She is currently teaching the Principles of Data Science and Analytics and Customer Behavior and Social Media Analytics courses at USM.